Microsoft crée sa propre puce pour son cloud

Microsoft a récemment annoncé la création de ses deux premières puces, Azure Maia 100 et Cobalt 100, conçues spécifiquement pour son infrastructure cloud Azure. Ces puces marquent une étape importante pour l'entreprise, qui cherche à réduire les coûts liés à l'utilisation des GPU de Nvidia, largement utilisés pour l'entraînement de modèles de langage lourds.

La puce Azure Maia 100 est une unité d'accélération d'IA conçue pour exécuter des charges de travail d'IA dans le cloud, notamment l'entraînement de grands modèles de langage. Elle sera utilisée pour alimenter certaines des charges de travail d'IA les plus importantes de Microsoft sur Azure, y compris le partenariat avec OpenAI. Cette collaboration avec OpenAI a impliqué des phases de conception et de test de la puce Maia. La puce est fabriquée sur la base d'un processeur TSMC de 5 nanomètres, avec 105 milliards de transistors, environ 30% de moins que le concurrent de Nvidia, l'AMD MI300X.

D'autre part, le processeur Azure Cobalt est une puce à 128 cœurs construite sur la base de la conception CSS Arm Neoverse et personnalisée pour Microsoft. Conçu pour les charges de travail en cloud, il offre des performances jusqu'à 40% supérieures à celles des serveurs Arm commerciaux, selon Rani Borkar, responsable des systèmes matériels et de l'infrastructure Azure chez Microsoft. Le processeur Cobalt est actuellement testé sur des charges de travail telles que Microsoft Teams, avec des plans pour mettre des machines virtuelles à la disposition des clients dans un avenir proche.

Ces deux nouvelles puces sont développées en interne par Microsoft, et leur conception est associée à une refonte approfondie de l'ensemble de ses serveurs pour optimiser les performances, la puissance et le coût. L'entreprise vise à répondre à la demande croissante en 2024, résultant de la forte augmentation de la demande pour les GPU H100 de Nvidia en 2023. Certains de ces GPU ont même atteint des prix de plus de 40 000 dollars sur eBay en raison de leur utilisation répandue dans l'entraînement de modèles de langage et d'outils d'images génératives.

Le processeur Cobalt et la puce Maia devraient être disponibles en 2024. Microsoft prévoit de déployer le processeur Cobalt pour une variété de charges de travail cloud, tandis que la puce Maia alimentera des charges de travail d'IA importantes sur Azure. Bien que les spécifications exactes et les performances détaillées ne soient pas encore divulguées, Microsoft est optimiste quant aux avantages concurrentiels que ces nouvelles puces apporteront à son infrastructure cloud.